Journal of Process Control 46 (2016) 75–83

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Distributed monitoring for large-scale processes based on multivariate statistical analysis and Bayesian method

Qingchao Jiang, Biao Huang*

Department of Chemical and Materials Engineering, University of Alberta, Edmonton, AB, T6G-2V4, Canada

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form 22 June 2016
Accepted 25 August 2016
Available online 31 August 2016

Keywords:
Distributed monitoring
Multivariate statistical analysis
Large-scale process
Bayesian method

ABSTRACT

Large-scale plant-wide processes have become more common and monitoring of such processes is imperative. This work focuses on establishing a distributed monitoring scheme incorporating multivariate statistical analysis and Bayesian method for large-scale plant-wide processes. First, the necessity of distributed monitoring is demonstrated by theoretical analysis on the impact of process decomposition on multivariate statistical process monitoring performance. Second, a stochastic optimization algorithm-based performance-driven process decomposition method is proposed which aims to achieve the best possible monitoring performance from process decomposition aspect. Based on the obtained sub-blocks, local monitors are established to characterize local process behaviors, and then a Bayesian fault diagnosis system is established to identify the underlying process status of the entire process. The proposed distributed monitoring scheme is applied on a numerical example and the Tennessee Eastman benchmark process. Comparison results to some state-of-the-art methods indicate the efficiency and feasibility.

1. Introduction

With increasing demands in plant safety and product quality, process monitoring is gaining considerable attention in both academic research and industrial applications [1–6]. Nowadays, the plant-wide processes are usually characterized by large-scale, multiple operation units and complex interactions, and monitoring of such processes has become an important issue [7,8]. In the same time, because of the wide use of sensor networks and distributed control systems, process data has become abundant. Data-driven especially multivariate statistical process monitoring (MSPM) methods progress rapidly.

Among the extensively researched MSPM methods, principle component analysis (PCA) and partial least square (PLS) serve as the most fundamental techniques and are generally used for monitoring processes with Gaussian distribution [9–12]. To deal with process dynamics, dynamic PCA and dynamic PLS methods which consider both the auto- and cross- correlations have been developed [13,14]. To deal with process non-Gaussianity, independent component analysis (ICA)-based methods have also been developed [15,16]. To deal with process non-linearity, kernel methods, such as kernel PCA (KPCA) [17–19], kernel PLS (KPLS) [20,21], and kernel ICA (KICA) [22], support vector dominant description (SVDD) [23,24] are generally employed. Other MSPM methods such as Fisher discriminant analysis (FDA) [25,26], canonical variate analysis (CVA) [1,27], etc., have also been used [6]. Although numerous applications of the MSPM have been reported, these methods may not function well for a large-scale process due to the centralized monitoring structure. The disadvantages of using the centralized monitoring in a large-scale process can be analyzed from two aspects:

(i) Practical considerations: First, fault-tolerance ability. A centralized monitor looks after all units of a system simultaneously. Once a fault in a unit is detected, the centralized monitor may limit its ability to detect further faults from other units that occur simultaneously. Second, reliability. A centralized monitor uses all measured variables in the computation. Once one variable is unavailable (such as delay, slow-sample rate) or one communication channel is blocked, the whole monitoring system may stop function. Third, economical efficiency. When there are geographical distributions of processes, for example, long distance between process units, it is natural for each unit to be equipped by a separate monitor.
(ii) Theoretical justification: Literature [28] provides a geometric explanation on the monitoring performance of the centralized and distributed PCA process monitoring. It has been shown that incorporating all variables into one centralized monitor could

* Corresponding author.
E-mail address: biao.huang@ualberta.ca (B. Huang).

http://dx.doi.org/10.1016/j.jprocont.2016.08.006
0959-1524/

actually degrade the monitoring performance. This paper will analyze the impact of process decomposition on the monitoring performance within the statistical analysis framework, proving that the monitoring performance could be improved by separating rather than incorporating all variables into one monitor.

Given the efficiency in dealing with large-scale processes, the multi-block or distributed monitoring scheme has gained significant attention. In multi-block or distributed monitoring methods, the process decomposition is one of the key steps, which can affect the monitoring performance significantly. Traditional multi-block monitoring schemes usually assume a process has been appropriately decomposed [29–31], however, in practical applications, process decomposition is a difficult task and remains an open question. Recently, the data-driven distributed monitoring methods have gained a particular interest [19,28,32]. Literature [28] developed a fault-relevant performance-driven variable selection method to achieve process decomposition; however, the number of local monitors would be significantly large as the number of fault increases. This paper will introduce how to perform performance-driven process decomposition with a limited number of local monitors.

Another issue in distributed monitoring is the fault diagnosis, which aims to identify the process status of the entire process. As one of the most widely applied probabilistic inference technologies, Bayesian method has shown its efficiency for various monitoring and diagnosis purposes. A framework of Bayesian diagnosis has been established by Huang [33], and within the framework, several researches have been conducted [34–37]. Recently, a Bayesian fault diagnosis system using optimal principal components (PCs) as evidence sources has been established in Ref. [34]. This study will introduce how the Bayesian diagnosis system can be applied in distributed process monitoring. The main contributions and novelty of this paper are summarized as follows:

(i) The impact of process decomposition on the monitoring performance is analyzed within the statistical framework of hypothesis testing, enhancing the distributed monitoring theoretical foundation.
(ii) Based on the theoretical analysis, a distributed monitoring framework is established, providing guidelines for designing a distributed monitoring scheme for large-scale processes.
(iii) Within the framework, a distributed monitoring scheme incorporating multivariate statistical analysis and Bayesian diagnosis system is developed. A performance-driven process decomposition method with a limited number of sub-blocks is proposed, facilitating the practical application feasibility.

The reminder of this article is organized as follows: In Section 2, the basic fault detection problem is reviewed and the impact of process decomposition on monitoring performance is analyzed. In Section 3, a data-based distributed monitoring framework is proposed and the developed distributed monitoring scheme is presented in detail. Then the developed distributed monitoring scheme is applied on a numerical example and the Tennessee Eastman (TE) benchmark process in Section 4. In Section 5, conclusions and future perspectives are presented.

2. Problem formulation

2.1. Basic statistical fault detection problem

The statistical fault detection tries to find out if there exists a fault in a process based on the statistical framework of hypothesis testing. Given a measurement model

$$\boldsymbol{x} = \boldsymbol{x}_N + \Theta f \tag{1}$$

where $\boldsymbol{x}_N \sim N(0, \boldsymbol{\Sigma})$ denotes the process data from normal status; f denotes a fault and the fault characteristics (i.e., magnitude and direction) are defined by the fault parameter Θ. Then the hypothesis testing can be formulated as: H_0, null hypothesis: $f = 0$, fault-free; H_1, alternative hypothesis: $f \neq 0$, faulty. A statistic is generally constructed to determine whether a measurement support the rejection of the null hypothesis. Let J denote a statistic and J_{th} be the threshold of the statistic, the probability $prob\left\{J > J_{th} | f = 0\right\}$ is called false alarm rate (FAR, or significant level α); the probability $prob\left\{J < J_{th} | f \neq 0\right\}$ is called non-detection rate (NDR).

The Hotelling T^2 statistic is generally constructed for fault detection purpose as [2]

$$T^2 = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}, \tag{2}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix. It has been theoretically proven that the Hotelling T^2 statistic provides the minimal NDR for a given FAR [2]. Given an acceptable FAR rate α, the threshold of the T^2 statistic can be determined by [2]

$$T_\alpha^2 = \chi_\alpha^2(m), \tag{3}$$

where $\chi_\alpha^2(m)$ denotes the chi-squared distribution with m degrees of freedom and significant level α. In some condition, the covariance matrix $\boldsymbol{\Sigma}$ may have numerical instability or with unfilled rank, and then singular value decomposition (SVD) on the covariance matrix is performed as [2]

$$\boldsymbol{\Sigma} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}^T. \tag{4}$$

Dividing the eigenvalue matrix $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_{pc} & 0 \\ 0 & \boldsymbol{\Lambda}_{res} \end{bmatrix}$ and the loading matrix $\boldsymbol{P}$ as $\boldsymbol{P} = \left[\boldsymbol{P}_{pc} \boldsymbol{P}_{res}\right]$ derives the PCA monitoring, in which two statistics are constructed as [2]

$$T_{PCA}^2 = \boldsymbol{x}^T \boldsymbol{P}_{pc} \boldsymbol{\Lambda}_{pc}^{-1} \boldsymbol{P}_{pc}^T \boldsymbol{x}, \tag{5}$$

$$Q = \| \left(\boldsymbol{I} - \boldsymbol{P}_{pc} \boldsymbol{P}_{pc}^T\right) \boldsymbol{x} \|_E^2 = \boldsymbol{x}^T \left(\boldsymbol{I} - \boldsymbol{P}_{pc} \boldsymbol{P}_{pc}^T\right) \boldsymbol{x}. \tag{6}$$

where $\boldsymbol{I}$ denotes the identity matrix. The thresholds of the statistics can be determined by $T_{PCA,th}^2 = \chi_\alpha^2(l)$ and $Q_{th} = \theta_1 \left(\frac{c_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right)^{1/h_0}$, where $l \leq m$ is the number of retained PCs in the dominant subspace, $\theta_i = \sum_{j=l+1}^{m} \left(\sigma_j^2\right)^i$, $i = 1, 2, 3$, $h_0 = 1 - \frac{2\theta_1 \theta_3}{3\theta_2^2}$, c_α is the normal deviate corresponding to the upper $1 - \alpha$ percentile for a given significance level α, and σ_j^2 is the j-th eigenvalue in $\boldsymbol{\Lambda}$.

2.2. Impact of process decomposition on monitoring performance

Given an acceptable FAR, the Hotelling T^2 provides the best monitoring performance for a multivariable process, i.e., the NDR are minimized [2]. However, for a large-scale process, the number of measured variables is generally large, and incorporating all measured variables in one monitoring model is not appropriate. In this